Universitat Politècnica de Catalunya

Universitat Rovira i Virgili
Universitat de Barcelona

Facultat d'Informàtica de Barcelona

Campus Nord Building B6

C/Jordi Girona, 1-3

Barcelona, Spain

08034

Master Thesis

Multiresolution community detection in Multilayer Networks

Bárbara Garza

18.06.2018

Supervised by

Sergio Gómez

Assistant Supervisor

Joan Matamalas

Acknowledgments

First of all I would like to thank my advisors Sergio and Joan... for their kind and continuous help, encouragement and guidance and for being so patient and supportive and for all their knowledge.

Furthermore I would like to thank my friend Sara for all her pep talks.

Abstract

Over the last few years, data everywhere has been growing and growing, we live in the era of Big Data. Because of the exponential growth of data, researchers, engineers, mathematicians and everyone who wants to make the best of this data, came up with several methods to categorize it and take advantage of it. Networks, specially multilayer networks, can be a robust representation of data from the simplest to the most complex depending on the information one wants to exploit. The WWW (World Wide Web), any social circle, neural networks, computational systems, a subway/metro system, are some examples of systems that can be represented as a network.

In recent years, Community detection has been very attractive in the field of complex networks; it provides a way to identify the substructures of o network that may be relevant in the studies of that given network [11] but nonetheless, community detection algorithms vary on their results because of the different methods and heuristics that are used to achieve the discovery of modules. Due to this differences, it is difficult to trust any algorithm; which is why it is important to identify several partitions of a network structure and see which one is more stable. We propose a multiresolution approach to an already good performer community detection algorithm called Infomap.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The aim of this thesis is to apply multiresolution to Infomap, a community detection algorithm based in the map equation that work in monolayer and multilayer networks. In this chapter, I explain what the thesis is about, the motivation and the structure of the thesis. Then in other chapters, we go over why implementing multiresolution in community detection algorithms for multi-layer networks is an interesting and a beneficial effort.

1.1 Motivation

There are two main objectives in this work. The first one is to bring InfoMap, a community detection algorithm applied in Multilayer Networks and the AFG Multiresolution method [2] together. This combination will provide a glimpse to the substructures in a Multilayer Newtwork. The second objective will help determine the validity of the proposed method by comparing the results to other community detection algorithms and normal Infomap in monolayer and multilayer networks as well as performing a layer by layer analysis of the multilayer networks. This thesis describes an approach to combining Multilayer Networks and a Multiresolution method in order to get more insights of a complex network.

1.2 Outline

This thesis is separated into 6 chapters.

Chapter 2 gives some overview of the fundamentals and background talked about in this thesis.

Chapter 3 describes Multilayer Networks, and the state of the art in community detection in these new kind of networks.

Chapter 4 describes the model proposed and what it tries to achieve along with the comparison measure we will be using against the original Infomap method.

Chapter 5 contains the experiments made and evaluation results based on well known metrics and benchmarks network for mono and multilayer networks.

Chapter 6 summarizes the thesis, describes the problems that occurred and gives an outlook about future work.

Chapter 2

Community Structure

This chapter is intended to give an introduction about relevant technologies and methods in the complex networks field regarding community detection. The sections are by no means written with the aim of providing a complete overview but sufficient background is provided so you can get the idea of what the thesis is trying to achieve by mixing all the following technologies together.

2.1 Community Detection in Complex Networks

Complex networks theory is a modern field of research that is spreading in many disciplines [5] such as biology, sociology, epidemiology, economics, transportation, physics, engineering, among others. It's emerging had to do with the ability of complex networks to represent systems as graphs of nodes and links being entities and their respective interactions or relationships. Complex networks distinguish themselves from random graphs because of their structure and because they model real systems.

The importance of community detection emerges from the natural structure in networks. As seen in real life, society offers an extensive range of group organizations as well as many networked systems exist in different areas of science. The information that the community structure reveals can be very valuable for scientists and for their respective studies.

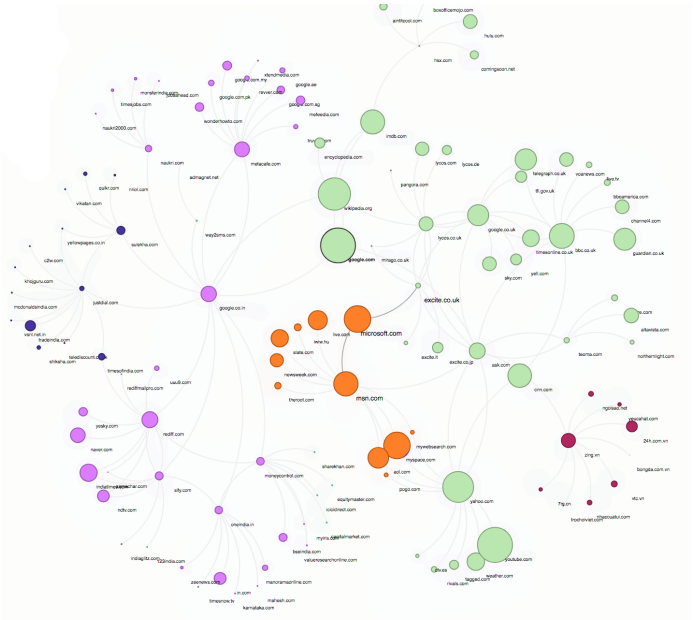
2.1.1 Networks

But what are networks? First, here we define a few key concepts used in the complex networks lingo and in this paper.

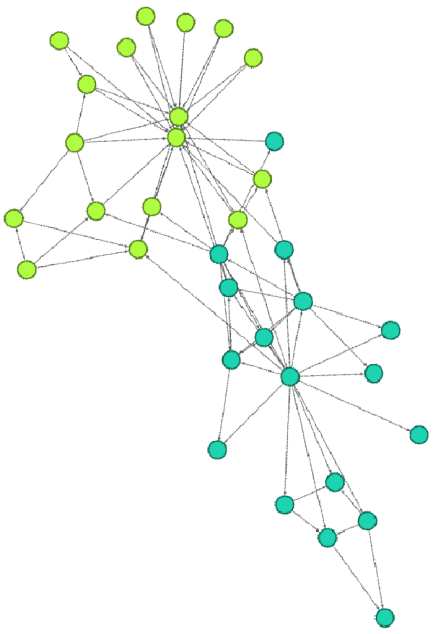
- Nodes: these represent the fundamental units of a network and are used to represent any entity with its respective attributes.
- Edges: these represent the links or relationships that connect nodes in different characteristic ways.
- Degree is the number of edges connected to certain node.

A network is a set of connected objects called nodes; these nodes are connected by edges that represent their relationship. In figure 2.1 we can see two examples of real networks, the first one is a very big and complex network and the second one a small one. Figure 2.1a represents the World Wide Web (WWW) Network, it shows key connections and communities of similar websites. The map includes over 1 million commonly visited websites. The size of the nodes is associated with a website popularity and importance. Figure 2.1b represents the Zachary's karate club network [29] which model consists on 34 members of and their relationships outside the club. You can observe two different colors in the network which means the two communities that appeared because of a conflict between the administrator and the instructor that led to a split of the club. The colors in both networks represent the communities detected of each one, more of community detection in the following section.

Networks can be different types depending on its topology. A network topology refers to the way the network is connected. It can variate depending on the edges connectivity between nodes. One way to classify the edges is their direction, the simplest connection can be undirected but they can also be directed. A simple example to explain directed networks can be a network of a Christmas raffle, everyone in family gathering is supposed to give 5 presents to 5 different family member, I might give my sister a present but she may not give one back to me. The second way to



(a) The WWW Network



(b) Zachary's karate club network

Figure 2.1: Examples of real networks

Figure 2.1a shows the internet network, its key connections and communities of similar websites. And 2.1b shows the Zachary Karate Club network, a network that represents the friendships between members of a karate club at a University

classify edges is their weight, an edge can be weighted or unweighted. An example of an unweighted network can be your Facebook social network, you can't measure the friendship, you just know that you are connected you your contact. And a book example of a weighted network can be a network of cities where the weight can be the distance between them.

Erdös-Rényi Random Graphs

The Erdös-Rényi random graphs are a simple type of networks where N number of nodes are connected to another with given the probability p. Many properties of these kind of networks are solvable in the limit of large graphs with the model having a Poisson degree distribution written as [23]:

$$p_k = \binom{N}{k} p^k (1-p)^{n-k}$$

The expected structure of a random graph depends on the probability of the parameter p which means the connection probability.

2.1.2 Community Detection

Having an extensive amount of data is not useful if you cannot do anything with it, that is one of the reasons why community detection has become very popular and important in network science; It gives insights of the modular structure of a network. Communities exists when groups of nodes have higher probability of being connected to each other instead of being connected to other nodes of the network. The previously mentioned means that nodes interact more strongly with the members of their community than they do with nodes of other communities in the event that they do at all [12].

Today, a very large amount of clustering algorithms exist to identify community structure. The most popular of the methods relies on modularity optimization but it was found that optimizing modularity fails to identify smaller communities. This

problem is known as resolution limit [11] described in more detail in the next section, 2.1.5.

Many community detection algorithms exist but only a few popular ones are often used. Among the most known algorithms we can find the following:

- Traditional methods like Hierarchical clustering
- Divisive methods like the algorithm of Girvan and Newman
- Modularity optimization methods
- Dynamic algorithms
- Methods based on statistical inference like block models
- Methods that find overlapping communities
- Multiresolution methods, etc

These and more algorithms are explained in [10]. In this work we will be mainly talk about community detection methods based on modularity like the Louvain method and the AFG multiresolution method. Furthermore we will talk about Infomap, a method based on information compression and flow [28].

2.1.3 Modularity

Given that detecting community structure is essential to clarify function and structure in complex networks, a crucial quantitative measure was introduced by Newman and Girvan [25] called modularity. Modularity was proposed for evaluatiing the "goodness" of a partition by comparing the number of links inside a given community or module with the expected value of a random network of the same size and degree sequence [11] also known as null model. Modularity [24] can be written as:

$$Q = \frac{1}{2w} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(w_{ij} - \frac{w_i w_j}{2w} \right) \delta(C_i, C_j) \tag{2.1}$$

where C_i stands for the community to which node i is assigned and the delta function $\delta(C_i, C_j)$ returns the value 1 if the nodes i and j belong in the same module and 0 alternatively. And

$$w_i = \sum_{j=1}^{N} w_{ij} \tag{2.2}$$

being the weighted adjacency matrix that represents the value of the weight in the link that connects the nodes i and j.

$$2w = \sum_{i=1}^{N} w_i = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \tag{2.3}$$

so $2w$ represents the total strength. With that in mind, the larger the modularity, the denser are the connections between nodes within a module and sparse connections between nodes belonging to other modules.

It is proven that maximization of modularity is an NP-hard problem [4], researchers have tried to optimize maximization with different heuristics but most of them end up falling in the same problem with another local maxima.

Louvain method

One method that optimizes modularity and is fast at unfolding communities in large networks better known as the Louvain method [3]. The method unfolds a complete hierarchical community structure, giving access to different resolutions of the network.

The technique works on a local optimization of the modularity in the neighborhood of each node. This way, a partition is identified and communities are replaced by supernodes. This same procedure is iterated until the modularity stops increasing. An advantage of this method is that it offers a fair settlement between the accuracy of the estimate of the modularity maximum and the computational complexity, which is linear.

The reason behind the choice to mention this method was its good performance and speed in the Lancichinetti and Fortunato comparative analysis of community detection methods [19]. This way we can compare two different approaches to community detection furthermore in our experiments.

2.1.4 InfoMAP

Infomap is an algorithm that detects communities in large networks by optimizing the map equation [27]. In order to detect communities, Infomap uses random walkers to explore the network and encodes the flows of the walks revealing important structures of the network. Each encoded flow is a module or community of the network and the modules with the minimum description length describe the optimal ones.

In comparison with other modularity methods, the map equation does not follow this measurement, it goes literally with the flow of the network and because of this reason the two methods can output very different results for the same networks.

To compress the random walk, the map equation benefits from the use of a Huffman code [15]; they abide optimally efficient symbol-by-symbol encoding and save space by assigning short codewords to the most common occurrences of events or objects and a larger one to the least common. This code applied to the nodes in a network, each codeword identifies a particular node, and the length of the codewords are obtained from a reasonably large random walk visits. Additionally to the codewords, the map equation uses a two-level description of the random walk by having a codebook, each codeword in the cookbook indicates the region the path is going, and this way the codewords can be reused in each region. The modules are obtained from the regions a random walker enters, usually when the walker gets to a region, it stays there for a while. When a region has a long persistence time, it gets its own codebook. An example of a random walk is represented in Figure 2.2.

Providing a two-level option, the random walker description decreases on average a 32% for this network. As a result of its information compression and flow approach, the Infomap algotithm was the best performer in the comparative analysis of clustering done by Lancichinetti and Fortunato on the LFR benchmarks [19].

2.1.5 Problems in community detection

Community detection is a difficult task and the problem has not been satisfactorily solved [10] even though several methods have been successful at finding communities.

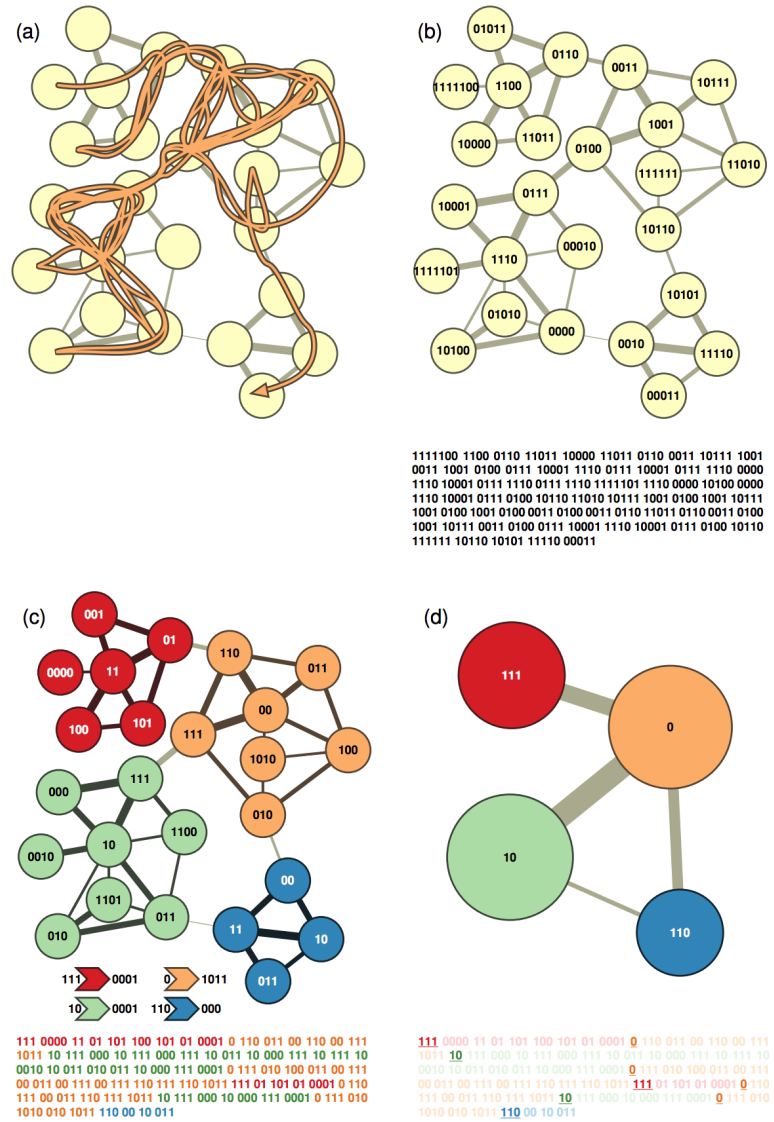


Figure 2.2: Representation of a random walk in InfoMap [27].

a) The orange line illustrates the trajectory of the random walker. b) Shows an efficient one-level description using Huffman codewords with the respective trajectory from a) in black. c) Shows a two-level description of the walk with its codebook that distinct modules or regions. The path below starts with the code from the codebook and everytime the walker switch regions it mentions the module code again so codewords for nodes are not to be confused. In d) we can see the codes that represent each module.

Some of the most common problems found is the controversy of how many clus-

ters are enough; other problems, arise even in accepted existing methods to detect communities.

A very noted problem is the one called resolution limit. Resolution limit appears when working with modularity optimization. Having a network with high modularity means that it has modules with dense connection within its nodes, and dispersed connections to other nodes of other modules. Nonetheless, modularity optimization, suffers a resolution limit, meaning that the algorithms implementing this approach fail to identify smaller communities show by Fortunato and Barthelemy [11].

Clara Granell explains the problem in a good and uncomplicated way in [6]: Imagine an image with a real size elephant and an ant, to see the ant clearly, we have to zoom in the image so much that the elephant practically disintegrates into pixels. Only a part of the elephant is noticeable when focusing on the ant.

The resolution limit comes from the very definition of modularity [10]. The null model defined in modularity, assumes that every node can get linked to any other node in the network, however, this assumption is far from reality. It is more reasonable to assume that each node interacts with just a limited part of the network. This signifies that the expected number of links between two modules decreases if the size of the network increases. Then, if a network is sufficiently large, the expected number of edges between two modules in the null model may be smaller than one; if this happens, one edge between two modules may be understood as a strong relationship between this two modules and modularity could merge these two no matter what. This is why, optimizing modularity in large networks would fail to find small communities even if they are well defined or even cliques.

Other methods have their respective problems too. Infomap, even though it seems to look unaffected by resolution limit [19], bears a problem known as the field-of-view limit, which means that communities tend to be over partitioned. The phenomenon develops because structural methods like Infomap contain an implicit scale and can only detect communities that spread within a range of effective sizes and might miss groups out of that range.

Having these problems, resolution limit or field-of-view limit, raises some concerns

about the reliability of communities detected by using a modularity optimization technique or Infomap because of the fact that this issue could have a large impact in practical applications. Because of this fact, multiresolution methods were proposed to try to deal with this issues.

2.1.6 The AFG Multiresolution Method

Multiresolution methods emerged by trying to tackle the resolution limit problem that occurs when optimizing the modularity in large networks. As mentioned before, small communities would get lost even if they are well defined.

With the aim of addressing the resolution limit issue, the AFG multiresolution method was introduced in [2] by Arenas, Fernandez and Gomez, still using modularity, but adding a parameter that controls the resistance of the nodes in order to form communities. The idea was that the community analysis was to be performed at multiple resolution scales. This screening of the topological structure allows us to see the most stable partitions and if any scale is more important than others by representing the community structure better.

The mathematical formula can be written as:

$$Q_{AFG}[w_{ij}, C, r] = Q[w_{ij} + r\delta_{ij}, C] \tag{2.4}$$

where the r represents the resistance, the parameter regulating the resolution of the partitions to find and $w_{ij} + r\delta_{ij}$ represents the new matrix from the original network with self-loops of weight r for every node. This way, screening through the scales allows to identify the number of communities per scale and what nodes form these communities.

The criteria for the scanning method was to scan from the macroscale, a single community with all the nodes to the microscale, where every node is its own community. By adding a self-loop node to every node, the internal strength of each node will increase allowing nodes to be isolated.

In this work, Arenas, Fernandez and Gomez tested their method with synthetic

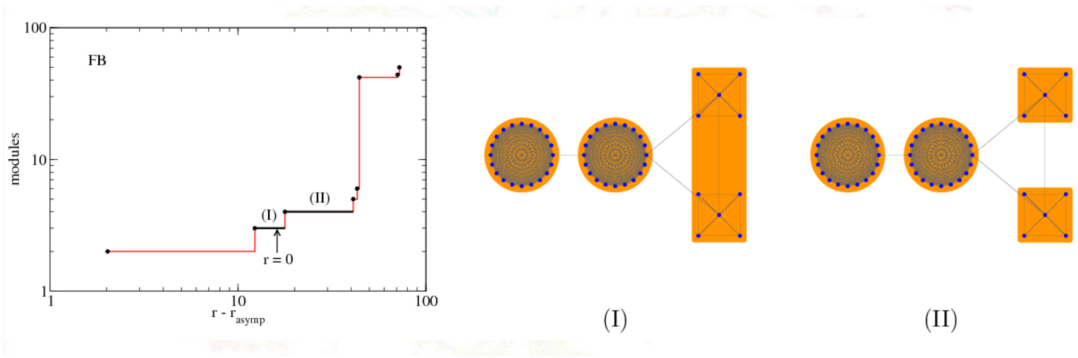


Figure 2.3: Fortunato & Barthelemy proposed network ran in the AFG method [2].

This network was proposed in [11] and the goal was to detect 4 communities as shown in (II), whereas (I) was normally detected when maximizing modularity, this was, having the two cliques of 20 nodes as two communities and the two small cliques of 5 nodes joined together. We can see that (II) is more stable thanks to the sweeping of the network in AFG.

and real networks. For the synthetic ones their results, the most stable partitions, corresponded to the predefined structures established a priori. For the real networks, the results correspond to previous knowledge in regards to the networks; with real networks, imposing certain number of communities can be difficult, the structure is indicated by analyzing the facts known about the network.

In a later work by Lancichinetti and Fortunato [20], it was proven that the methods dedicated to bypass the resolution limit had in fact a resolution limit. One of those methods was the AFG method. The limitation of the method is a split-and-join problem [], which still merges small cliques and splits large cliques. Split-and-join can not be solved by changing the resolution, the problem still lies in modularity as this measurement is not appropriate in the event that communities of very different sizes coexist.

The AFG method failed in detecting correct number of communities in a proposed benchmark that consisted in a Erdös-Rényi (ER) network of 400 nodes with a degree of 100 connected to two cliques of 13 nodes each, the two cliques shared a link between them. The aim was to get three communities but the method failed and detected only two, the 400 one and a 26 community. Later, we will use this benchmark to

compare our work with other methods.

Chapter 3

Multilayer Networks

As we know, networks can illustrate several types of relationships between objects or entities. In spite of that, trying to express multiple properties in a single link can convey in loss of valuable information [16]. This is why having multiple layers of connectivity is important to enhance the understanding of a complex system where in one single network, the nodes can exhibit different relationships simultaneously [8].

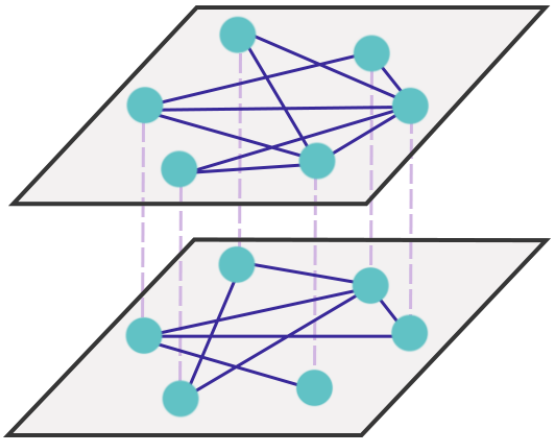


Figure 3.1: Multiplex network

A type of multilayer network where the nodes in one layer must be represented in all of its layers.

A good example of a multilayer network can be a social network where individuals

have ties to other individuals and those relationships are of different types, family, work, school, romantical, etc. Having just a unique type of relationship will be far from reality.

Another real example, in this case, of a multiplex network is the network of airport routes, represented in figure 3.2. In this network the nodes represent all the airports and the relationship represented in each layer is the airline connecting these nodes. Of course, this can be represented in a monoplex network but in a multilayer environment, it has a way better representation and visualization.

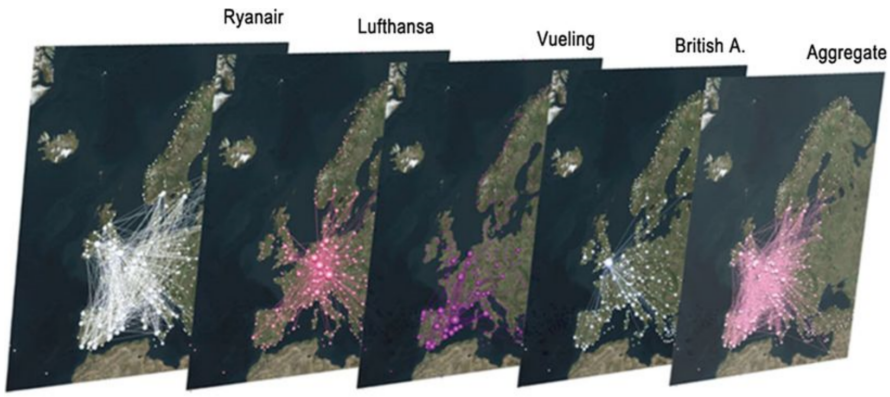


Figure 3.2: European airports network, where each layer represents a different airline [8].

Multilayer networks are the most reality based type of networks. In summary, a multilayer or multiplex network are several layers of complex networks and from this network we want to have the ability to get the same information and metrics that we are able to from single layered or monoplex networks: for this reason, and due to the fact of its aggregated complexity we are going to use this kind of networks to prove our work.

3.1 Community detection in multilayer networks

In spite of the great advances that have been made in the detection of communities in monolayer networks, the advances in community detection of multi-layer networks

is very limited due to their added complexity.

It has been very complicated to tackle community detection but some have tried generalized monolayer algorithms, but of course, carrying with it the problems that existed in detecting partitions in monolayer networks to begin with. However, researchers have developed methods despite the complications their method could bring. For example, Mucha et al [22], proposed a multiplex model for describing a similar multirelational network and developed a generalized framework of network quality functions as modularity in order to study the structure of multislice networks or multilayer networks. Other methods exists

As for modularity maximization methods, they are already an NP-hard problem, adding more layers means adding more complexity

One obstacle in community detection is the difficulty of validating the communities. This issue comes from the definition of community itself, because even if has statistical significance it may not make sense with the knowledge base one has from the data. There exists several perceptions from a network and the difficulty increases when the network increases and of course, in this case, when the layers start to stack up. Even if the community detection algorithms are tested with synthetic networks, it is hard to recreate reality in the synthetic and using real networks can also be an arduous job because of the a priory work and study a network has to have.

We choose Infomap among other community detection algorithm because of the fact that is not a method that relies in modularity and that gives us an advantage over other methods.

Chapter 4

Multiresolution in Multilayer Networks

This section describes the model we used to carry out our experiments explained in the next chapter, chapter 5. The methods used are the ones described in the previous chapter and we will introduce two more methods to which we compared our results.

4.1 Model

The model that we propose consists of using the Infomap algorithm described in 2.1.4 but applying a resistance parameter or self-loop simulating the idea of the AFG method discussed in 2.1.6 with the goal in mind of getting the structure that best defines the desired networks.

The intention is to get all the possible scales in monolayer as well as for multilayer networks to identify which partitions are the more stable according to the network. This goal can be achieved by increasing the probability of a node to teleport to itself, this way the analysis will start with a single community containing all of the nodes (the macroscale) in the network and furthermore, it will divide until each node becomes its own community (the microscale). This approach will make Infomap multiresolution.

As we discussed before, the ability to detect communities can be of significant

importance in the studies of a network, because of the fact that community detection provides insight into how networks function and how the topology affects the relationship between the elements in the network. In multilayer networks, each layer enhances the understanding of the system the network is trying to represent; providing the best, most fitting and stable communities will only enrich and avert possible information loss.

4.2 Normalized Mutual Information

In order to measure and compare our results, we will be using a known similarity measure, NMI; approved and very often used in tests for community detection algorithms . This measure of similarity of communities was borrowed from information theory, which proved to be reliable [21]. This symmetric measure quiantifies the statistical common information between two clusters and can be written as:

$$I_{norm}(X,Y) = \frac{2I(X,Y)}{H(X) + H(Y)'} \tag{4.1}$$

where $I(X,Y)$ denotes the mutual information between X and Y and $H(X)$ and $H(Y)$ stand for the entropy of X and Y respectively. A normalized version of the equation 4.1 ranges from 0 to 1 where it is equals 1 if the partitions are identical and is equal or close to 0 when the partitions are very different or independent.

Chapter 5

Experiments and Results

This chapter covers the resolution of the performed experiments and its respective results along with their discussion. First, our approach was tested in monolayer networks, synthetic and real ones. After, it was tested in multilayer networks, also in some synthetic and real ones. As earlier explained, in real networks, the results are a bit more difficult to determine because of the fact that nothing from the topology hints the presence of a more relevant structure in a given network; the corroboration comes after a structure is found with the known facts and meaning of the network.

We will talk about the comparison of our results with other methods and with the same Infomap algorithm without a selfloop using NMI measurement when applicable.

5.1 Monolayer testing

In this section we show the results of the method and furthermore we compare it to 'normal' Infomap results and the obtained results from running the networks with the Louvain method. For the experiments, we took into account some examples of synthetic and real complex networks. And for some of the networks, we can compare the results with the actual real partitions and for some others only with the results obtained from other methods.

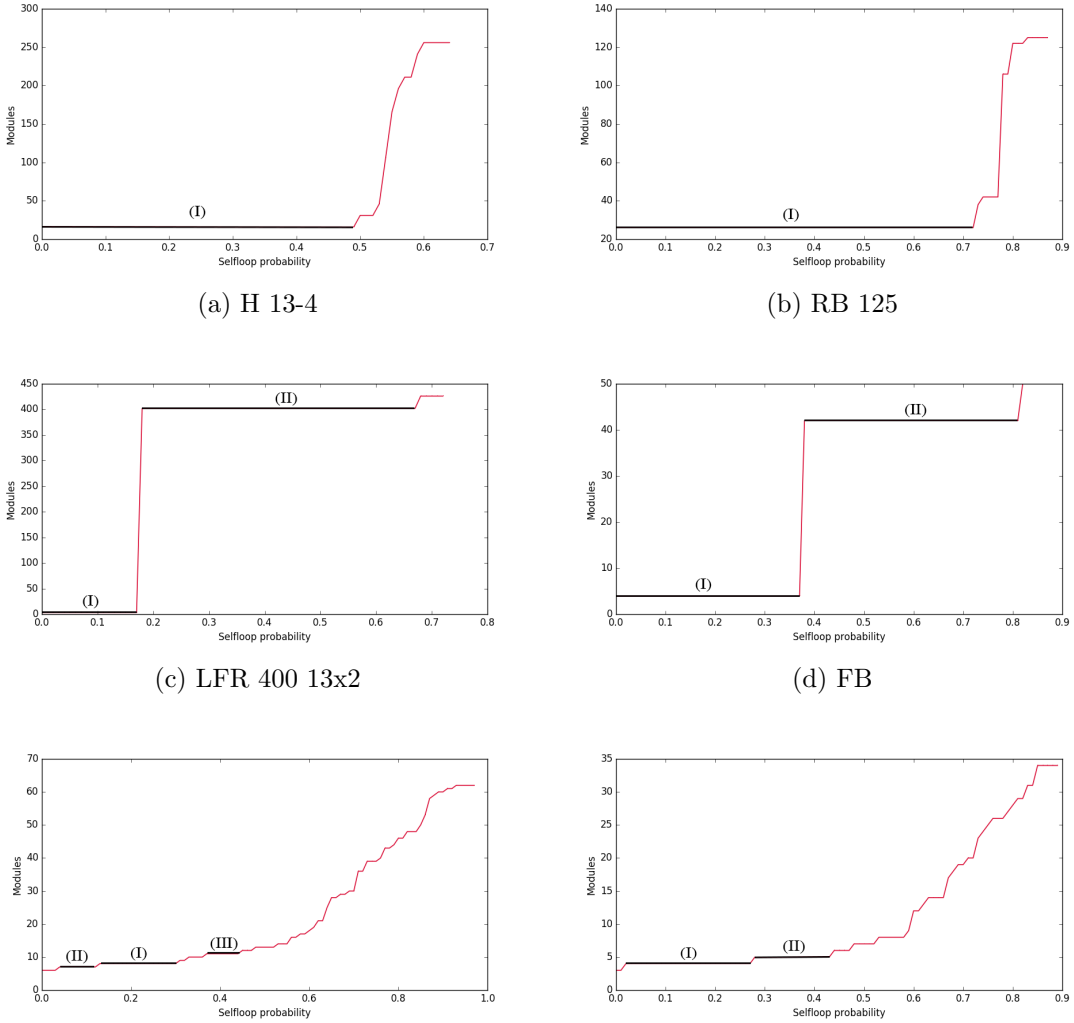
The networks analyzed here and also in [2] are the following:

- The H 13-4 network [1] corresponds to a homogeneous in degree network that has two hierarchical levels. It has 256 nodes and the first level has 16 communities of 16 nodes each and the second one being the external community is formed of four groups of 64 nodes.
- RB 125, proposed by Ravasz and Barabasi in [26], is a hierarchical scale-free network.
- A LFR A 400 random graph with two 13 node cliques, proposed by Lancichinetti and Fortunato in [20] with this network, a resolution limit was proven in the AFG method as mentioned in [6].
- The FB network, propsed by Fortunato and Bartélemy in [11] aimed to demonstrate the resolution limit of modularity. The network is formed of four cliques, two of 20 nodes each and two of 5.
- Dolphins network, a directed social network of bottlenose dolphins. Where each node represents a dolphins of a bottlenose dolphin community living off Doubtful Sound, a fjord in New Zealand. An edge indicates a frequent association. The dolphins were observed between 1994 and 2001 [17].
- Zachary Karate Club Network [29], as mentioned in chapter 2, is a network that represents the friendships between members of a karate club at a University. The network is formed by 34 nodes and 78 edges.

For statistics, each experiment was run 50 times, each one with 100 resistance values ranging from 1 to 0, were this number means the probability for a node to teleport to itself.

In table 5.1 we can see the summary of the experiments run for monolayer networks. As we can see, Applying a selfloop to Infomap, the most stable number of modules is still equivalent to the number of modules the normal Infomap computes.

In figure 5.1 a) we can see the most stable partitions by screening the network. For a), that represents the H 13-4 network, we see that the most stable partition and practically the only one, is for 16 modules; the method fails to detect another hierarchical level which is formed by 4 modules of 64 nodes each as shown in figure 5.2.



(a) H 13-4

(b) RB 125

(c) LFR 400 13x2

(d) FB

(e) Dolphins

(f) Zachary

Figure 5.1: Multiresolution in monolayer networks

These are the results of running Infomap with a selfloop parameter. Applying these, we can screen the network starting with a probability of teleporting to itself of 0 to 1 ranging from the closer it can be to the macroscale to the microscale

Table 5.1: Summary of experiments in monolayer networks

		Infomap with selfloop	**Infomap (normal)**	**Louvain Method**		
Network	**Real**			**1**	**0.75**	**0.5**
H 13-4	4 or 16	16	16	5	4	4
R 125	5 or 25 or 26	26	26	11	5	7
400 13 13	3	3	3	7	2	2
FB	4	4	4	4	3	2
Dolphins	2	8	6	5	4	2
Zachary	2	4	3	4	3	2

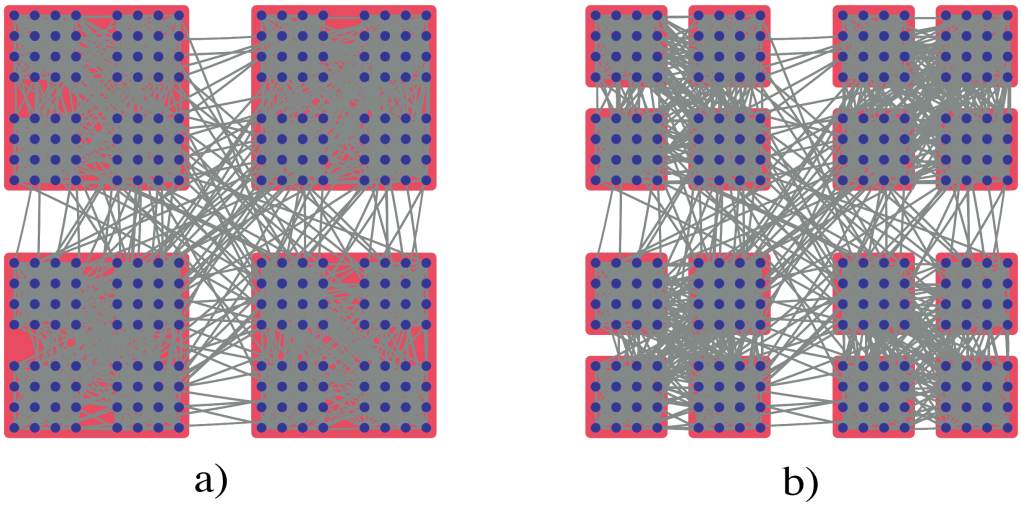


Figure 5.2: H 13-4 network with its possible partitions [2]

For b), the RB 125 network, there is clearly one partition that stands above the small one, the more stable partition points out 26 communities; for this network, another two partitions had to be identified and Infomap also failed in that job. A 5 community structure and a 25 one as shown in figure 5.3.

In c) we can see that the partition that looks more stable is not the one we are looking for, which shows 402 communities, this is the big clique of 400 nodes already

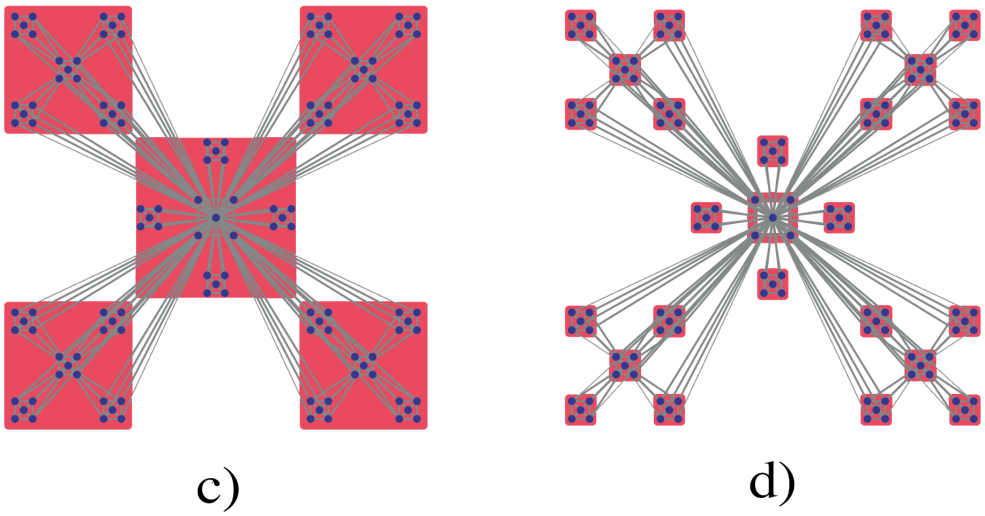


Figure 5.3: RB 125 network with its possible partitions [2]

disconnected and the two nodes belonging to the 13 nodes clique. Thanks to previous knowledge of the network we can make a qualitative decision of picking the next stable partition which is the one with 3 communities as shown in figure 5.4.

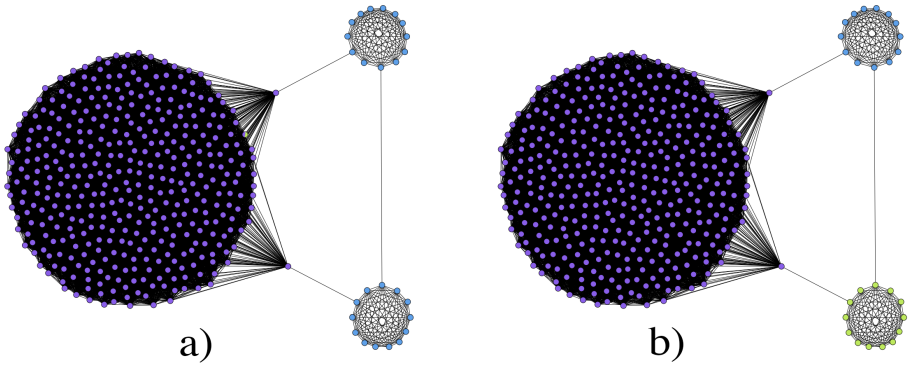


Figure 5.4: 400+13+13 network with its possible partitions [6]

In a) you can see two partitions, one with 400 nodes and one with the two cliques of 13 nodes each and in b) you can see the expected number of partitions which is 3. One for the 400 ER network and two for the two small cliques.

In d), the FB network, the first stable partition is the one with 4 communities,

and it never gets 3 partitions as modularity methods do and as shown in figure 5.5. The second partition when all the 20 nodes cliques become their own community while the other two stay linked.

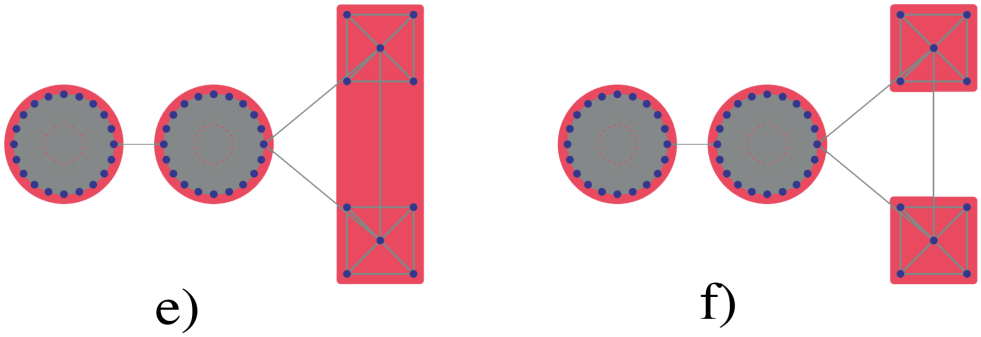


Figure 5.5: FB network with its possible partitions [2]

For the dolphins network in e), we can detect three stable partitions but it doesn't match the real structure of the network as seen in 5.1, Infomap with selfloop and the normal one fail and the minimum number of communities we get is 6. For the final experiment in monolayer networks, we can see that the algorithm also fails detecting the number of communities in the Zachary network, being the stabler partition with 4 communities and the minimum detected 3.

Some of the previous networks, where also tested by Arenas, Fernandez and Gómez in the AFG [2], but in contradiction to our method, the AFG method was more successful in finding all the stable partitions defined a priori for the networks picked.

5.2 Multilayer testing

In this section we show the results of the multiresolution Infomap algorithm applied to multilayer networks and we compare it to 'normal' Infomap results. Furthermore, we analyze each layer of each network as a monolayer network. For the experiments, we took into account some examples of synthetic and real complex networks:

The networks analyzed are the following:

- Star Wars social network, a toy network assembled by Evelina Gabasoba [13] and available in muxViz [8]. This social network is composed of the relationships of the members of the StarWars universe from Episodes 1 through 6. The network is divided into layers by episode.
- The London Transport System [9] is a network formed of nodes representing the train stations in London and edges encode the routes between stations. Each layer is for Underground, Overground and DLR respectively.
- A 2x2 network, with overlapping communities. It is a very small network of four layers of 8 nodes each; each layer designed to have two communities, each one of 4 nodes and they rotate in each layer.
- Pierre Auger Collaboration, a group of theoretical and experimental scientists working at the Pierre Auger Observatory. The collaborators work in different research topics and they can work in several topics at the same time.
- The LFR 500, is a network of 500 nodes generated with the LFR benchmark described in [18], with different mixing parameter to add complexity to the network.

As with the monolayer testing, each experiment in this section was run 50 times, each one with 100 resistance values ranging from 1 to 0 where this means as above, the probability for a node to teleport to itself. In the table 5.2 we can see the results obtained from our proposed approach, the stabler partition in Infomap with multiresolution and the normal output from 50 runs in Infomap. We use NMI, the normalized mutual information explained in section 4.2 to compare both partitions from the radatools package in [14].

First, we start with the analysis of the StarWars network, showin in Figure 5.6, and as we can see in table 5.2, it is one of the networks that gives us the mos different result between our approach and the normal Infomap algorithm. As for layer-by-layer analysis, the results from both types of Infomap and the Louvain method with different resolutions is not very contrasting.

Table 5.2: Summary of experiments in multilayer networks

Networks	**Infomap with selfloop**	**Infomap (normal)**	**NMI**
Star Wars	20	15	0.8352
London Transport System	63	60	0.9824
2x2	4	4	1.0
Pierre Auger Collaboration	108	108	0.99
LFR 500 0.25	323	319	0.9545
LFR 500 0.50	339	345	0.9552
LFR 500 0.75	361	361	1.00
LFR 500 1.00	364	353	0.953

The next network is the London Transport System shown in Figure 5.7, in which the overall comparison is not much different with a NMI of 0.9824, the most stable partition with out algorithm is very similar to normal Infomap. As for the analysis by layer, normal Infomap and Infomap with selfloop are also very similar in contrast to Louvain method. Nonetheless, because I know what this network is about, public transportation, I would trust more Infomap results because this algorithm is based on flow, and public transportation is connected and meant for someone to walk on, follow it; the nature of Infomap algorithm based on random walkers encourages us to choose it.

Unlike the other real networks, the 2x2 benchmark shown in Figure 5.8 had well defined communities in each layer but it did not help our case as all the Louvine method did not had any issue finding the communities either. This just proves that it does not matter the size of the network in relation to finding the number of communities like when our approach failed at detecting the Zachary Club communities in previous section.

The results in the Pierre Auguer Collaboration network were also very similar according not only to the number of communities found by both algorithms but also

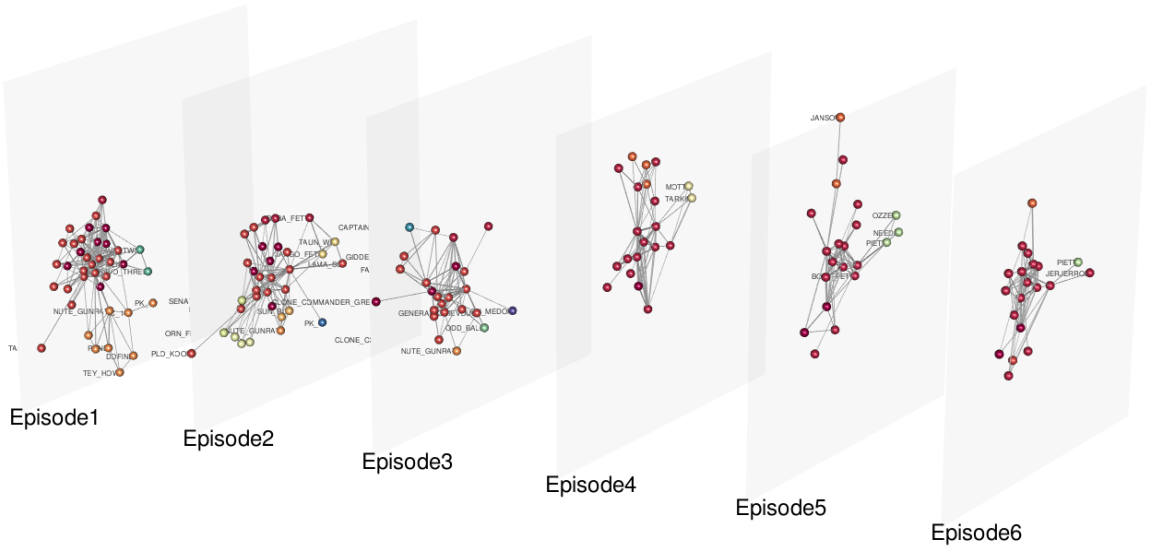


Figure 5.6: StarWars multiayer network [8].

in the analysis of layers. This must mean that the number of communities found must be close to the ground truth. For this case, it is challenging to select other partitions that are stable. If we inspect Figure 5.10, we can appreciate that there is no visible stable partition, they seem to change in every increasing step of the selfloop probability

For the last network, the LFR benchmark with 500 nodes, we have 4 different structures depending on the mixing parameter. We generated bechmarks with mixing parameter of 0.25, 0.5, 0.75 and 1 respectively and the NMI for the partitions obtained from both Infomap versions for each one is very close to 1.

5.3 Overall results

As seen in the tables and figures, we did not get very different results from out self-loop approach to normal Infomap, even for the layer by layer and monolayer testing. For monolayer network analysis, even though our approach was successful in finding the expected number of modules in almost all the networks, when the network had two or more stable and accepted partitions not all were found in comparison with the

Table 5.3: Star Wars Network analysis by layer

	Infomap with selfloop	**Infomap (normal)**	**Louvain Method**		
Layer			**1**	**0.75**	**0.5**
Episode 1	59	58	59	58	56
Episode 2	63	63	63	63	61
Episode 3	70	69	71	70	69
Episode 4	74	74	75	74	74
Episode 5	74	74	74	74	74
Episode 6	74	74	75	75	74

Table 5.4: London Transportation Network analysis by layer

	Infomap with selfloop	**Infomap (normal)**	**Louvain Method**		
Layer			**1**	**0.75**	**0.5**
Tube	137	135	116	113	111
Overground	302	302	296	294	292
DLR	336	333	329	329	329

results in the AFG method [2]. For some networks as the H 13-4, the RB 125, and Zachary networks it fails to detect the partitions with less communities, this can be in correlation to the field-of-vew problem talked about in section 6.2. So despite the scanning of the mesoscale of the network, the Infomap algorithm not always starts at one community that involves all of the nodes in the network, it over partitions since the very start.

As for multilayer results, the success of the algorithm is even more difficult to prove as discussed in section 2.1.2. The lack of more benchmarks accepted and available for testing and for real networks to have communties already categorized makes the task very difficult to achieve. What helps our case here is the analysis by layer, the

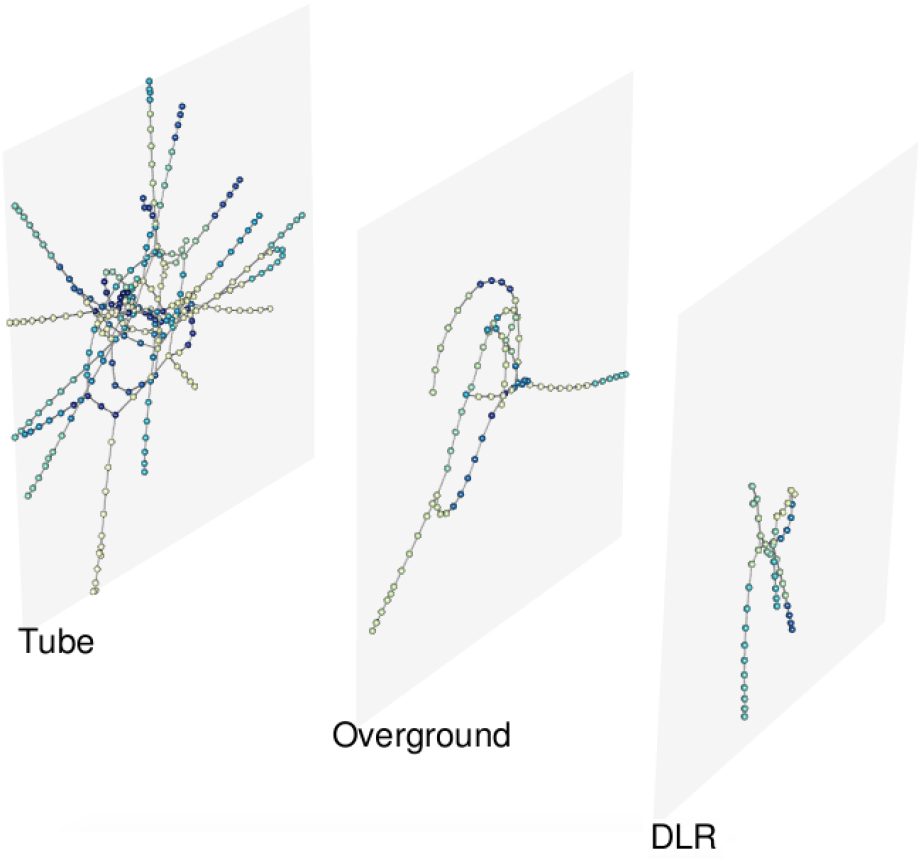


Figure 5.7: London Transportation multiayer network [8].

results of Infomap with and without self loop are very similar to the ones obtained in different resolutions with the Louvain method based on modularity.

But despite of the not great results, it is still very important to screen the network and using a self-loop is a very good approach that still needs more tweaks regarding other parameters within Infomap. Analyzing all the possible scales a network can have gives so much more insight to researchers and developers so they can decide based of previous knowledge and facts about the network which partition is the one that fits best.

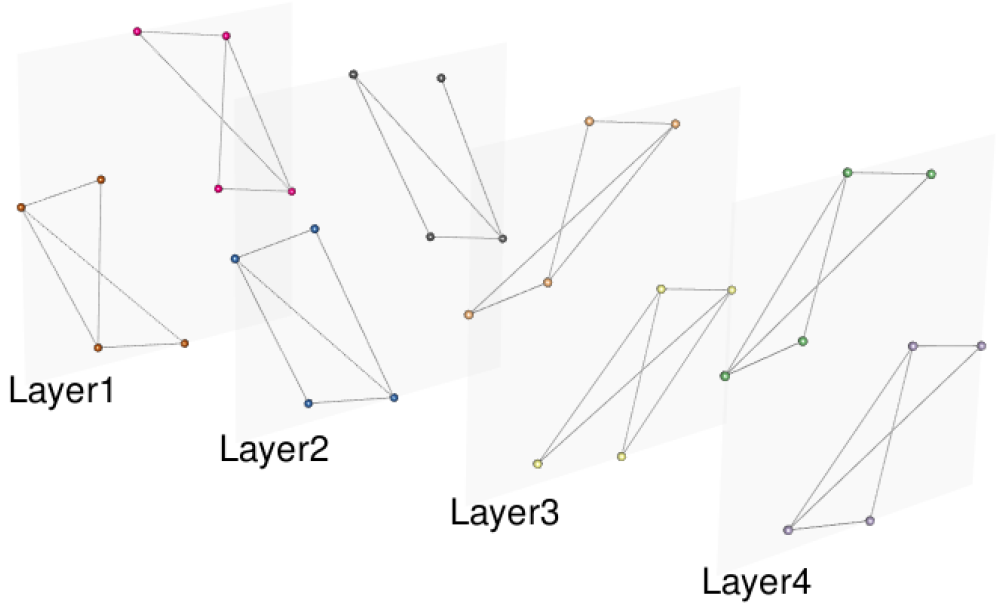


Figure 5.8: 2x2 benchmark multiayer network [8].

Table 5.5: 2x2 Network analysis by layer

	Infomap with selfloop	**Infomap (normal)**	**Louvain Method**		
Layer			**1**	**0.75**	**0.5**
Layer 1	2	2	2	2	2
Layer 2	2	2	2	2	2
Layer 3	2	2	2	2	2
Layer 4	2	2	2	2	2

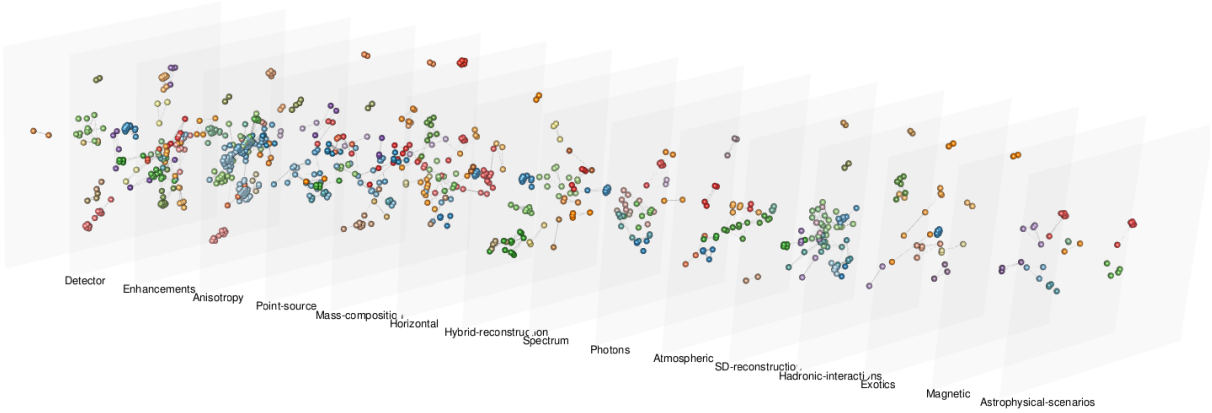


Figure 5.9: Pierre Auguer Collaboration multiayer network [8].

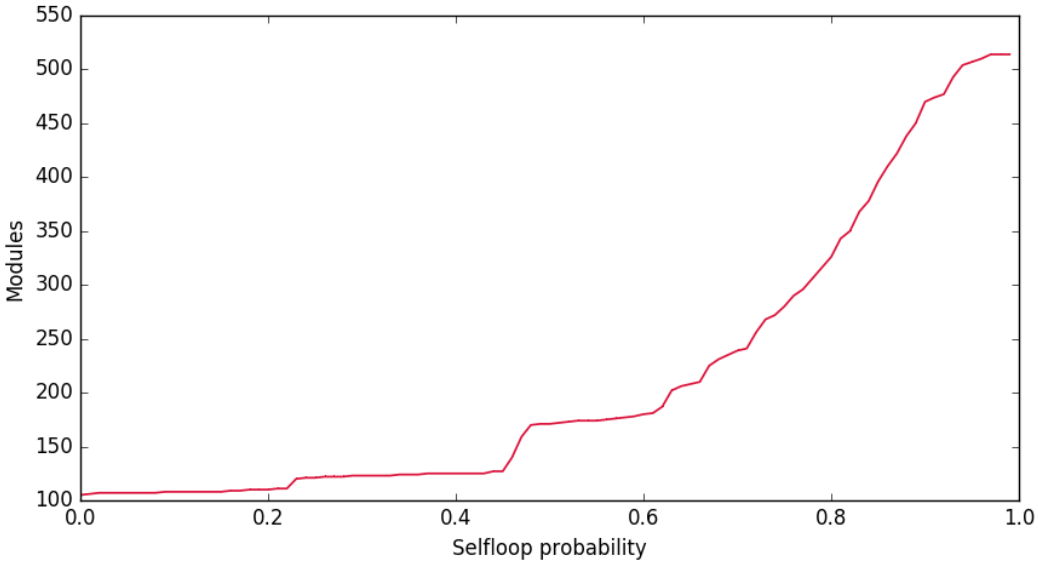


Figure 5.10: Pierre Auguer Collaboration mesoscale analysis [8].

Table 5.6: Pierre Auguer Collaboration Network analysis by layer

	Infomap with selfloop	**Infomap (normal)**	**Louvain Method**		
Layer			**1**	**0.75**	**0.5**
Neutrinos	498	496	497	497	496
Detector	376	376	370	370	369
Enhancements	297	297	291	290	298
Anisotropy	479	478	477	477	477
Point-source	462	461	461	460	460
Mass-composition	444	444	441	441	441
Horizontal	495	495	494	494	494
Hybrid-reconstruction	460	459	456	456	455
Spectrum	480	479	479	479	479
Photons	503	503	503	503	503
Atmospheric	494	494	493	493	492
SD-reconstruction	453	452	450	449	449
Hadronic-interactions	495	495	495	494	494
Exotics	504	504	504	504	504
Magnetic	495	495	495	495	495
Astrophysical-scenarios	506	506	506	506	506

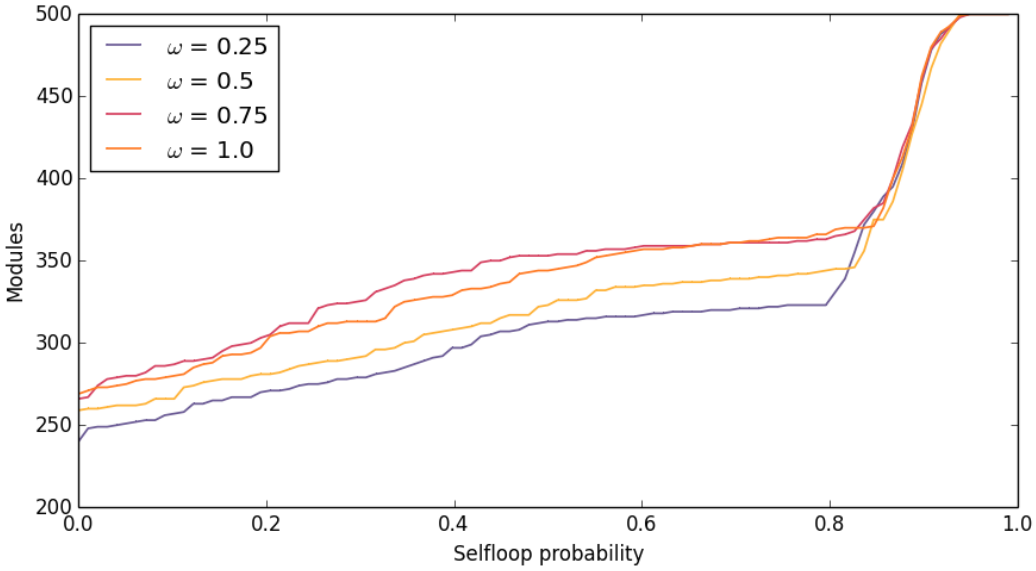


Figure 5.11: LFR 500 multiayer networks with mixing parameter.

Chapter 6

Conclusion

In this work, we presented an multiresolution approach for the comunity detection Infomap algorithm. we have tested the proposed algorithm in monolayers as well as multilayers and for the later, we have applied it once again for each layer. Even though the results are not as different as applying normal Infomap algorithm to retrieve the communities of a network, our approach scanning multiple resolutions in a network is successful due to the extra information available of the structure of the network by scanning the mesoscale. This way, several other possible partitions can be detected that are not in normal Infomap. Another positive outlook for these experiments is to reaffirm that Infomap, even with its field-of-view limit is a stable algorithm which has proven to output the most stable partition of a network in the mayority of experiments.

6.1 Summary

The work can be summarized in these steps:

- Analysis of the state of the art in complex networks, community detection and multiresolution methods. All the technologies described in chapter 2 and 3
- Selection of relevant benchmarks for the analysis of the implementation and the comparison method.

- Implementation of code that screens monolayer and multilayer networks using Infomap with selfloop.
- Automatization of Infomap runs for testing.
- Dataset converter and also creating benchmarks from scratch when no source was available.
- Evaluation of the proposed solution by running each experiment 50 times for monolayer and multilayer networks.

6.2 Problems Encountered

Some of the major problems that were encountered while developing the project include:

Variety in the lack of standardization in multilayer networks lingo. One of the intentions of [16] is to help standardize aspects of multilayer networks; it seems like every researcher and developer structures datasets and algorithms inputs however they please, so everyone has their own way. This issue is quite time consuming when looking for alternatives for comparing methods or testing their algorithms and data.

Other problem was that almost no one had their code available and when it was available, it seemed like it was no longer maintained, bad documented and full of execution errors.

I also had difficulties performing the evaluation due to the lack of benchmarking tools available for community detection on multilayer networks, it seemed to be that the only one available was the LFR benchmark.

6.3 Future Work

For future work, I will enhance the way the script is executed, now it has some tweaks and is not as straight forward as I would like. Several steps need to be made to get all the information needed, not everything is automated as well as for the convertion of the data. Even though there is a script for converting datasets gathered by and

available from Domenicos website [7], Infomap has one way of input, Pajek other and muxViz another; and some of those convertions where done manually for each dataset.

The repository also needs to be well documented and public, to achieve the last one, everything that is not mine needs to be cited. A good objective would be to add more multilayer community detection algorithms to get a full overview of the structure of a network and make a qualitative decision about the correct number of modules a network has but this can be a goal for the long future because of the problem discussed in 6.2, some of the few algorithms that tackle multilayer networks, the code is not available and even if I try my best I am no mathematician and implementing them from scratch would be even a longer task. For this work it was out of the scope so It was not done.

Bibliography

[1] A. Arenas, A. Díaz-Guilera, and C. J. Pérez-Vicente. Synchronization reveals topological scales in complex networks. *Phys. Rev. Lett.*, 96:114102, Mar 2006.

[2] A. Arenas, A. Fernández, and S. Gómez. Analysis of the structure of complex networks at different resolution levels. *New J. Phys. 10, 053039*, 2008.

[3] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 10:10008, Oct. 2008.

[4] U. Brandes, D. Delling, M. Gaertler, R. Goerke, M. Hoefer, Z. Nikoloski, and D. Wagner. Maximizing Modularity is hard. *ArXiv Physics e-prints*, Aug. 2006.

[5] H. Cherifi. *Complex Networks and Their Applications*. 01 2014.

[6] A. A. Clara Granell, Sergio Gomez. Hierarchical multiresolution method to overcome the resolution limit in complex networks. *International Journal of Bifurcation and Chaos*, 22(7):1250171, 2012.

[7] M. De Domenico.

[8] M. De Domenico, M. A. Porter, and A. Arenas. Muxviz: a tool for multilayer analysis and visualization of networks. *Journal of Complex Networks*, 3(2):159–176, 2015.

[9] M. De Domenico, A. Solé-Ribalta, S. Gómez, and A. Arenas. Navigability of interconnected networks under random failures. *Proceedings of the National Academy of Sciences*, 2014.

[10] S. Fortunato. Community detection in graphs. *CoRR*, abs/0906.0612, 2009.

[11] S. Fortunato and M. Barthelemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, Jan. 2007.

[12] S. Fortunato and D. Hric. Community detection in networks: A user guide. *CoRR*, abs/1608.00163, 2016.

[13] E. Gabasova. StarWars Social Network. `https://github.com/evelinag/`, 2015. [Online; accessed June 2018].

[14] S. Gomez and A. Fernandez.

[15] D. A. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, Sept 1952.

[16] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter. Multilayer networks. *CoRR*, abs/1309.7233, 2013.

[17] KONECT.

[18] A. Lancichinetti and S. Fortunato. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *ArXiv e-prints*, Apr. 2009.

[19] A. Lancichinetti and S. Fortunato. Community detection algorithms: A comparative analysis. , 80(5):056117, Nov. 2009.

[20] A. Lancichinetti and S. Fortunato. Limits of modularity maximization in community detection. *CoRR*, abs/1107.1155, 2011.

[21] A. Lancichinetti, S. Fortunato, and J. Kertesz. Detecting the overlapping and hierarchical community structure of complex networks. *ArXiv e-prints*, Feb. 2008.

[22] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela. Community Structure in Time-Dependent, Multiscale, and Multiplex Networks. *Science*, 328:876, May 2010.

[23] M. E. J. Newman. Random graphs as models of networks. *eprint arXiv:cond-mat/0202208*, Feb. 2002.

[24] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Phys. Rev. E*, 69:066133, Jun 2004.

[25] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, Feb 2004.

[26] E. Ravasz and A.-L. Barabási. Hierarchical organization in complex networks. *Phys. Rev. E*, 67:026112, Feb 2003.

[27] M. Rosvall, D. Axelsson, and C. T. Bergstrom. The map equation. *European Physical Journal Special Topics*, 178:13–23, Nov. 2009.

[28] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. USA*, page 1118, 2008.

[29] W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473, 1977.